Predicting Toxicological Endpoints of Chemicals Using Quantitative Structure-Activity Relationships

Author(s): Paul Harten[1], Douglas Young[1], Todd Martin[1] and Raghuraman Venkatapathy[2]

Affiliation(s): [1]U.S. EPA, Office of Research and Development, National Risk Management Research Laboratory, Sustainable Technology Division, Cincinnati, OH
[2]Pegasus Technical Services, Inc., Cincinnati, OH

Introduction

Globally, the chemical industry and Regulatory Agency's such as the U.S. EPA spend millions of dollars in testing and assessing the health risks associated with chemicals. The risk management process is currently conducted using experimental data.

"Oh, go on then, you know me, I'll try anything once."

Problem

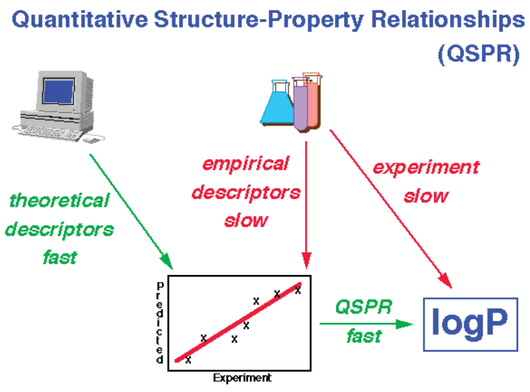
There are significant gaps in the availability of experimental toxicity data for most toxicological endpoints.

Solution

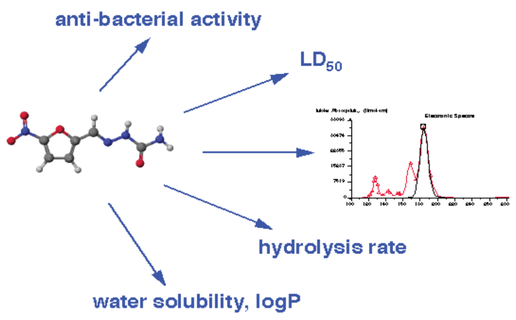
Use Quantitative Structure-Activity/Property Relationships (QSARs/QSPRs) to estimate toxicity

QSARs/QSPRs:

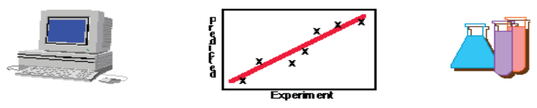
- Can relate biological activity (e.g., LOAEL, mutagenicity, LD_{50}) to physicochemical properties
- Properties depend on the structure of the chemical alone
- Properties can be calculated using computers



Structure-Property Relationship



QSAR & QSPR Benefits



- estimates almost any property or activity
- faster and cheaper than experiment
- can screen hypothetical compounds
- may provide some mechanistic insight

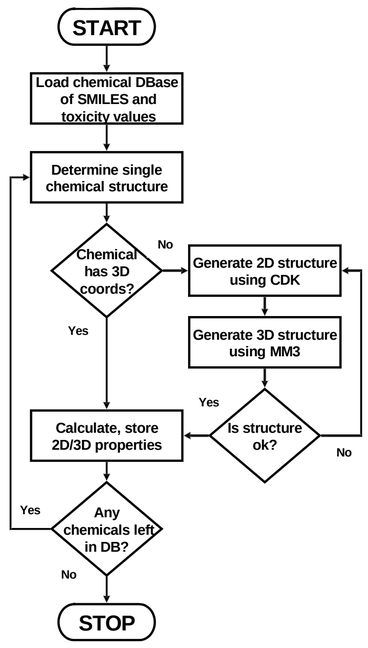
Abstract

Quantitative structure-activity relationships (QSARs) are being developed to predict the toxicological endpoints for untested chemicals similar in structure to chemicals that have known experimental toxicological data. Based on a very large number of predetermined descriptors, an analysis finding chemical clusters of similar descriptors will be performed utilizing Ward's method. For all chemical clusters that have a reasonable number of chemicals, a QSAR will be developed to predict a toxicological endpoint for untested chemicals of similar structure. Optimal combinations of the descriptors will be found by implementing a genetic algorithm that searches through the space of all predetermined descriptors to find what combinations of these descriptors produce the most accurate QSAR prediction. Each QSAR developed will have to meet certain statistical criteria to be considered valid. Because the search through descriptor space involves millions upon millions of calculations, this research will effectively utilize EPA's high-performance computing resources.

Phase I. Develop QSAR Equations

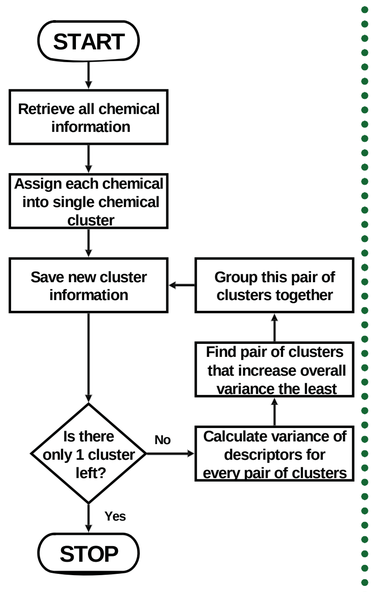
APPROACH: For each toxicity endpoint, the QSAR development phase of this project is divided into three parts: (A) Calculate descriptors, (B) Analyze chemical clusters and (C) Generate QSAR equations and validate using regression analysis and genetic algorithm.

A. Calculate descriptors



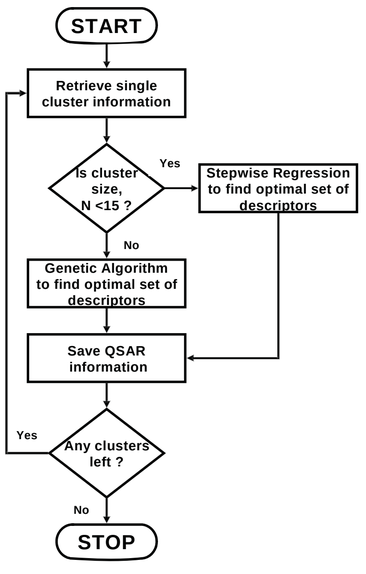
- Calculate 3-D coordinates of atoms using molecular mechanics (MM3), if unavailable
- Calculate approx. 800 2-D and 3-D descriptors
- 3-D descriptors are calculated using MOPAC

B. Analyze chemical clusters



- Group chemicals in database into clusters of chemicals with similar descriptor values
- All descriptors are used in the cluster analysis process
- Based on Ward's method
- Clusters contain between 3 and 250 chemicals

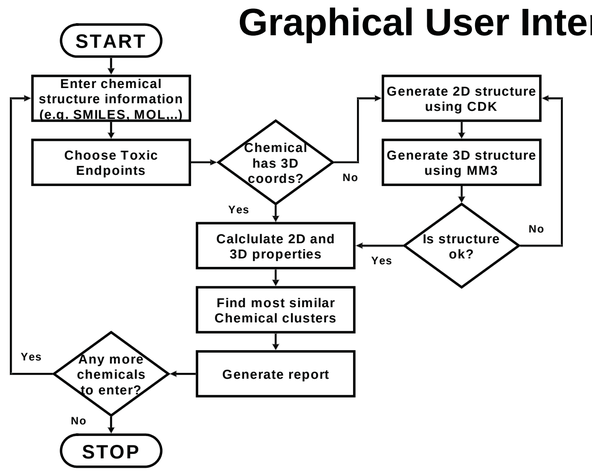
C. Generate QSAR equations



- Final QSAR equation generated using stepwise regression or genetic algorithm depending on number of chemicals in cluster
- Each cluster has a QSAR equation
- QSAR equations will be validated using internal (Leave-One-Out and Leave-Group-Out) and external validation datasets

Phase II. Estimate Chemical Toxicity

Graphical User Interface (GUI)



- Predicts toxicity using QSAR of most similar clusters
- Report provides 5 best predictions and associated confidence intervals
- Report also presents average toxicity prediction and range of predictions

epascienceforum 2006
Your Health • Your Environment • Your Future